OrdinalCLIP: Learning Rank Prompts for Language-Guided Ordinal Regression

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents a language-powered paradigm for ordinal regression. Existing methods usually treat each rank as a category and employ a set of weights to learn these concepts. These methods are easy to overfit and usually attain unsatisfactory performance as the learned concepts are mainly derived from the training set. Recent large pre-trained vision-language models like CLIP have shown impressive performance on various visual tasks. In this paper, we propose to learn the rank concepts from the rich semantic CLIP latent space. Specifically, we reformulate this task as an image-language matching problem with a contrastive objective, which regards labels as text and obtains a language prototype from a text encoder for each rank. While prompt engineering for CLIP is extremely time-consuming, we propose OrdinalCLIP, a differentiable prompting method for adapting CLIP for ordinal regression. OrdinalCLIP consists of learnable context tokens and learnable rank embeddings; The learnable rank embeddings are constructed by explicitly modeling numerical continuity, resulting in well-ordered, compact language prototypes in the CLIP space. Once learned, we can only save the language prototypes and discard the huge language model, resulting in zero additional computational overhead compared with the linear head counterpart. Experimental results show that our paradigm achieves competitive performance in general ordinal regression tasks, and gains improvements in few-shot and distribution shift settings for age estimation.

1 Introduction

Ordinal regression in computer vision aims to predict a rank number [6] or continue value [21] y for a given data $\boldsymbol{x}$. For example, age estimation aims to estimate the age of a given face image while image aesthetic assessment predicts the aesthetic score for an image. As a fundamental problem, ordinal regression has attracted increasing attention due to its broad applications [18, 27, 24, 15].

A simple solution for ordinal regression is to directly regress a scale value [35]. It is implemented with a single neuron and optimized with L_1 or L_2 loss. However, it suffers from limited performance. Popular approaches [36, 40, 35] usually discretize the continuous target space into different bins or treat the rank labels as different numbers. Then they regard these discrete values as different categories and perform the classification using a linear layer. These methods are essentially learning *a set of weights to model the concept of order*. There are two main challenges for these methods. First, treating ranks as independent class categories fails to grasp the ordinal property. Second, as the learned concepts are mainly derived from the training set, these approaches are prone to overfit and usually attain unsatisfactory performance.

Since learning the rank concept from the image domain alone is prone to overfitting, we can leverage multimodal information to alleviate this issue. The human language contains rich semantic information and prior knowledge. We consider simultaneously borrowing the rank concept from the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

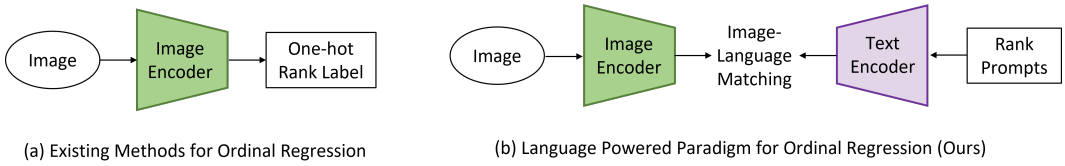


Figure 1: The key idea of our paradigm. Existing methods usually learn the rank concepts from the training data. Instead, we leverage the language priors to learn the rank concepts. We treat each rank category as text and extract the language prototypes with a well-learned text encoder. We align image features into the well-learned language latent space.

language domain. Specifically, each rank label is not only regarded as a class category, but also linked to a sentence describing the corresponding rank, such as "this person is 23 years old". In this way, our model not only learns the concept of ranks defined on the vision dataset, but also exploits the common knowledge of rank in the language domain. In this paper, we propose a language-powered paradigm for ordinal regression, which *learns a* ***language prototype*** *for each rank from a well-defined language latent space*. We aim to associate each rank category with its language concept to alleviate the overfitting issue. We use CLIP [33], which learns an image encoder and a text encoder with large-scale image-text pairs and achieves impressive performance on zero/few-shot knowledge transfer, to leverage the language priors. Specifically, we convert each rank label into text input (prompt) and employ the pre-trained giant text encoder in CLIP to extract language prototypes for all ranks. Then we reformulate ordinal regression as an image-text matching framework, where image embeddings are obtained from a learnable image encoder and a set of language prototypes are served as the text features to be matched. We aim to align the image features into the fixed language space. Since the prototypes are obtained from a fixed language model, we are somehow distilling the language knowledge from the CLIP model. Moreover, the prototypes are constrained in the well-learned language latent space, which is also a kind of regularization leading to stronger generalization.

Recent research [48] has found that the text input (prompt) plays a key role in the performance and choosing the right prompt is extremely time-consuming. Instead of using hard prompts, we propose to learn rank prompts for ordinal regression. The proposed OrdinalCLIP consists of learnable context tokens and learnable rank embeddings. Learnable context tokens are shared across all ranks and used to model the context words. As large language models usually do not fully learn the continuous property of the numbers [42], we also propose to learn rank embeddings by interpolating between several base features, which explicitly model the continuous nature of rank embeddings. The key idea of this paper is shown in Figure 1.

We summarize the main contributions of this paper as follows: (1) We present a new paradigm that learns language prototypes from a fixed language model for ordinal regression. To the best of our knowledge, our method is the first language-powered paradigm for ordinal regression. (2) We introduce learnable rank embeddings, which explicitly model the continuous nature of ranks to better exploit language priors from the text encoder in CLIP. (3) Comprehensive experiments illustrate the effectiveness of the paradigm. OrdinalCLIP achieves competitive performance on three real-world tasks, along with improvements in few-shot and distribution shift settings for age estimation.

2 Related Work

Ordinal Regression in Computer Vision: Given an input image, ordinal regression in computer vision aims to map the input to a rank or a continuous value. Many popular methods [35, 10] adopt a classification framework. For example, Rothe *et al.* [35] posed facial age estimation as a deep classification problem. Then they computed expected values over the softmax-normalized probabilities for inference. As the simple classification formulation treats different ranks as independent classes, some work [10, 28] attempts to model the relations between ranks. Geng *et al.* [10] proposed label distribution, which regards a sample as an instance associated with a specific label distribution. The mean-variance loss was introduced in [28] to leverage the label ambiguity. Chen *et al.* [3] proposed ranking-CNN with a series of basic CNNs to learn the rank. There are also some papers [11, 20, 12, 43, 14] modeling ordinal regression from the perspective of probability. Gustafsson *et al.* [11] proposed an energy-based model within a probabilistic regression formulation. Li *et al.* [20] introduced probabilistic ordinal embeddings to model the data uncertainty for regression.

DeepGUM [17] used a Gaussian-uniform mixture model to achieve robust performance for outliers. Pan *et al.* [29] presented self-paced deep regression forests to achieve more robust and less biased models. These methods still mainly learn the rank concept from the data, which makes them easy to overfit and leads to sub-optimal performance. Our paradigm aims to learn the concept from the language to alleviate the above issues.

Language Model: Recent years have witnessed the great success of pre-training general language representations [4, 2, 34, 32, 5]. ELMo [32] was proposed to extract context-sensitive features. BERT [5] employed masked language models to learn deep bidirectional representations. CLIP [33] proposed to learn representations from image-text pairs. The experimental results show that the CLIP is much more efficient at zero/few-shot transfer, which demonstrates the strong power of language. Encouraged by the excellent performance of CLIP, many CLIP-powered methods have been proposed. Patashnik *et al.* [30] proposed StyleCLIP, which leveraged the power of CLIP models to develop a text-based interface for StyleGAN. Wang *et al.* [44] introduced ActionCLIP for video action recognition. ActionCLIP follows a paradigm "pre-train, prompt and finetune", which achieves strong performance on action recognition tasks. Gao *et al.* [9] proposed the CLIP-Adapter to use the CLIP model on downstream tasks. Zhang *et al.* [47] further proposed Tip-Adapter to enhance the CLIP's few-shot capability without training. Wav2CLIP [46] was proposed to learn robust audio representations by distilling from CLIP models. In this paper, we leverage the language priors from the CLIP model. To better exploit the language priors from CLIP, we further propose differentiable rank prompts to model the continuous property of the numbers. The most relevant works are [48] and [49]. Zhou *et al.* [48] presented a new approach named context optimization. This method learns the context of prompts, which attains significant performance improvements. Zhou [49] also proposed an improved version of [48] by incorporating the visual features into the text features, which improves the generalizability of the model. Note that [48] only optimizes the context embeddings. Thus directly applying it to ordinal regression leads to degraded performance due to the lack of ordinal property. We do not compare with [49], as we want to solely investigate how language priors introduced as language prototypes could help ordinal regression tasks.

3 Proposed Approach

3.1 Problem Statement

The goal of ordinal regression in computer vision is to estimate the rank number or continue value for a given image. Although we can directly regress this value with a single neuron, a popular baseline is to regard it as a classification task and employ additional constraints trying to make use of the ordinal property. Mathematically, we denote the rank space as $\mathcal{R} = \{r_j | 0 \leq j < C\}$, where r_j denotes a rank category and C is the number of ranks categories. During training, we randomly sample a batch of images $\boldsymbol{X} = \{\boldsymbol{x}_i | 0 \leq i < B\}$ and the associated labels $\boldsymbol{Y} = \{\boldsymbol{y}_i | 0 \leq i < B\}$, where B represents the batch size and $\boldsymbol{y}_i \in \{0, 1\}^C$. The classification baseline treats each rank as a class number. Assuming the rank label for a sample $\boldsymbol{x}_i$ is r_j, we construct the one-hot label by setting $y_{i,l} = 1$ if $l = j$, otherwise $y_{i,l} = 0$, where l is an index and $0 \leq l < C$.

The images are first sent to an image encoder to extract image features $\boldsymbol{f} = \{\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{B-1}\}$, where $\boldsymbol{f}_i \in \mathbb{R}^d$ is the extracted features of image $\boldsymbol{x}_i$. We use d to denote the feature dimension. Then the image features are sent to a linear layer with a set of parameters $\boldsymbol{W} = [\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{C-1}]^\top \in \mathbb{R}^{C \times d}$ and $\boldsymbol{b} = [b_0, b_1, \ldots, b_{C-1}] \in \mathbb{R}^C$:

$$l_{i,j} = \boldsymbol{w}_j^\top \boldsymbol{f}_i + b_j \,, \tag{1}$$

where $l_{i,j}$ is the predicted logit for rank r_j of image $\boldsymbol{x}_i$. Then the model is optimized with a cross-entropy loss:

$$\mathcal{L}_{cls} = \frac{1}{B} \sum_{i=0}^{B-1} \sum_{j=0}^{C-1} -y_{i,j} \log \frac{\exp(l_{i,j})}{\sum_{j=0}^{C-1} \exp(l_{i,j})} \,. \tag{2}$$

Many existing methods develop additional technologies based on the classification formulation, such as label distribution [10, 8, 7], mean-variance loss [28], and so on. However, these methods simply treat each rank as a number and learn the rank concept mainly from the training data leading to the overfitting problem along with unsatisfactory performance. In this paper, we propose a language-powered paradigm, which learns the rank concept from language to address the above issues.

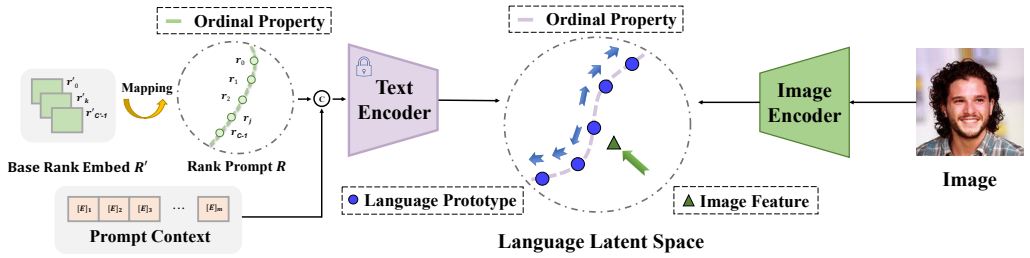


Figure 2: The framework of OrdinalCLIP. We regard rank categories as text and employ a language model to leverage the language priors. For each rank, we concatenate its word embedding and learnable prompt context. Then they are sent to a language model to extract the corresponding language prototype. To preserve the ordinal property of language prototypes, we explicitly construct the ordinal rank embeddings that are interpolated from several base rank embeddings. We found the ordinality of the rank embeddings can be implicitly propagated toward the language prototypes.

3.2 Language Prototypes

We believe that natural language contains rich prior knowledge. To preserve ordinality, we propose to link the learned rank category to its language concept. To leverage the language power, we employ the text encoder of CLIP [33] as the language model. CLIP follows a vision-language pre-training framework and learns representations from image-text pairs, resulting in a joint vision-language latent space. As the CLIP model shows strong performance in downstream tasks, we implement our method with the CLIP-powered text encoder.

We reformulate the ordinal regression task as an image-language matching problem, which consists of a text encoder and an image encoder. The text encoder is used to extract language features with the fixed CLIP language model. As the parameters of the text encoder are fixed, the text is mapped to a well-learned language latent space, which encodes rich language priors.

To leverage the language priors with the text encoder, we treat the rank categories as words. Then we obtain a set of words for the rank categories $\{[r_0], [r_1], \ldots, [r_{C-1}]\}$. To obtain the language embedding for each rank, we construct a sentence for these ranks and send the entire sentence to the text encoder. For example, we can construct a sentence of *"a person at the age of* $[r_j]$*"* for the rank $[r_j]$. Then each word of this sentence is mapped to a 512-D word embedding vector.[1] All the word embeddings are sent to the text encoder to obtain a text embedding. In this way, each rank category is mapped to a text embedding in the language latent space. We treat the mapped embedding as the prototype of its corresponding rank category. As these prototypes lie on the well-learned language manifold and encode the language concept, we name them as **language prototypes**. Formally, we use $\{\boldsymbol{p}_j \in \mathbb{R}^d | 0 \leq j < C-1\}$ to represent the ℓ_2 normalized language prototypes, where $\boldsymbol{p}_i$ corresponds to the language prototype of the rank r_j and d is the feature dimension.

The input images are processed by an image encoder to extract image features. We can employ any popular vision backbones as the image encoder. While the parameters of the text encoder are fixed, we train the entire image encoder to align image features to the language latent space. For a batch of images $\boldsymbol{X}$, we denote the ℓ_2 normalized image features as $\boldsymbol{I} = \{\boldsymbol{I}_i \in \mathbb{R}^d | 0 \leq i < B-1\}$, where d is the feature dimension. As an image-language matching framework, we calculate the similarity scores between the two modalities through the inner product and obtain a similarity matrix $\boldsymbol{A} = [a_{i,j}]$:

$$a_{i,j} = \boldsymbol{I}_i \cdot \boldsymbol{p}_j^\top \,, 0 \leq i < B-1, 0 \leq j < C-1 \,. \tag{3}$$

We use image-text contrastive loss to optimize the network, which consists of an image-to-text loss and a text-to-image loss following the practice of CLIP [33]. For image-to-text loss, we first normalize the rows of the matrix $\boldsymbol{A}$ using a softmax layer to obtain the matrix $\boldsymbol{A}' = [a'_{i,j}]$:

$$a'_{i,j} = \frac{\exp(a_{i,j}/T)}{\sum_{j=0}^{C-1} \exp(a_{i,j}/T)} \,, \tag{4}$$

[1] Or several 512-D sub-word embeddings due to the usage of BPE [38]

where T is the temperature parameter. We minimize the discrepancy between the matrix $\boldsymbol{A}'$ and label $\boldsymbol{Y}$ using KL divergence. For text-to-image loss, there may be zero or multiple hits. Therefore we construct the label matrix $\boldsymbol{Y}''$ by normalizing the non-zero columns of $\boldsymbol{Y}$ [44]. Besides, we obtain a column-normalized matrix $\boldsymbol{A}'' = [a''_{i,j}]$:

$$a''_{i,j} = \frac{\exp(a_{i,j}/T)}{\sum_{i=0}^{B-1} \exp(a_{i,j}/T)} \,. \tag{5}$$

We also employ a KL divergence to minimize the discrepancy between the matrix $\boldsymbol{A}''$ and the label matrix $\boldsymbol{Y}''$. In the end, we have the image-text contrastive loss:

$$\mathcal{L} = \frac{1}{2}[\frac{1}{B}\sum_{i=0}^{B-1} KL(\boldsymbol{Y}_{i,\cdot}, \boldsymbol{A}'_{i,\cdot}) + \frac{1}{C}\sum_{j=0}^{C-1} KL(\boldsymbol{Y}''_{\cdot,j}, \boldsymbol{A}''_{\cdot,j})] \,. \tag{6}$$

3.3 Learning Rank Prompts

As illustrated in [33, 48], different text input, which is known as prompt, may give very different performances. Instead of designing the prompt context, we directly learn m word embeddings $[E]_0, [E]_1, \ldots, [E]_{m-1}$ to serve as the prompt context, which is concatenated with the word embeddings of ranks and is sent to the text encoder.

To further strengthen the ordinality of our model and improve its performance, we propose to learn the rank embeddings to preserve the order of the language prototypes in the language latent space. The reason why we cannot directly apply CLIP to regression is that the human-built prompts from the rank labels lack ordinality (see Figure 3), which leads to degraded performance. Inspired by the arithmetic property in the word embeddings space [25, 31, 1], whose typical example is the following embedding arithmetic equation: "King $-$ Man $+$ Woman = Queen", we choose to maintain the order of rank embeddings to preserve the order of the language prototypes. By explicitly constructing continuous rank embeddings, we can also improve the ordinality of the language prototypes (also see Figure 3). We regard this ordinality as the natural outcome of continuity.

We derive the ordinal rank embeddings from a set of base rank embeddings via interpolation. Suppose we have a set of base rank embeddings $\boldsymbol{R}' = [\boldsymbol{r}'_0, \boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_{C'-1}]$, where C' denotes the number of the embeddings, and $\boldsymbol{R}' \in \mathbb{R}^{512\times C'}$. Note that in practice, C' should be a relatively small number compared with C (*i.e.*, $C' << C$). Then the j-th rank embedding could be obtained by:

$$\boldsymbol{r}_j = \sum_{k=0}^{C'-1} w'_{j,k} \cdot \boldsymbol{r}'_k \,. \tag{7}$$

where $0 \leq k < C'$, and $\boldsymbol{W'} = [w'_{j,k}] \in \mathbb{R}^{C\times C'}$ is the interpolation weights:

$$w'_{j,k} = \frac{I(w_{j,k})}{\sum_{k=0}^{C'-1} I(w_{j,k})} \,, w_{j,k} = |j - \frac{C-1}{C'-1} \cdot k| \,. \tag{8}$$

where $I(\cdot)$ is the interpolation function, and $|\cdot|$ outputs absolute value. We simply propose two interpolation approaches, which are *linear interpolation*: $I(w_{j,k}) = 1 - \frac{w_{j,k}}{C-1}$ and *inverse-proportion interpolation*: $I(w_{j,k}) = \frac{1}{w_{j,k}+\epsilon}$. We set ϵ to 1e-5 for numerical stability. Figure 2 presents the details of the rank prompt of OrdinalCLIP.

Although our method uses a huge language model to exploit the language concepts for training. We can only employ the learned language prototypes during testing. In this way, our method does not bring any additional computational and storage overhead compared with the linear head alternative.

4 Experiments

4.1 Age Estimation

Datasets: The task of facial age estimation is to predict the age of a given facial image. We test our method on the widely-used MORPH II [13] dataset and Adience [19] dataset. Morph II contains

Table 1: Results on MORPH II.

Methods	MAE	Year
AGEn [41]	2.52	2018
BridgeNet [21]	2.38	2019
AVDL [45]	2.37	2020
POE [20]	2.35	2021
CNN Baseline	2.63	-
Zero-shot CLIP [33]	14.45	2021
CoOp [48]	2.39	2021
OrdinalCLIP	**2.32**	-

Table 2: Results on the Adience face dataset.

Methods	Accuracy(%)	MAE
CNNPOR [22]	57.4 ± 5.8	0.55 ± 0.08
GP-DNNOR [23]	57.4 ± 5.5	0.54 ± 0.07
SORD [6]	59.6 ± 3.6	0.49 ± 0.05
POE [20]	60.5 ± 4.4	$\mathbf{0.47 \pm 0.06}$
CNN Baseline	56.0 ± 5.9	0.56 ± 0.09
Zero-shot CLIP [33]	25.4 ± 3.5	1.50 ± 0.20
CoOp [48]	60.6 ± 5.5	0.50 ± 0.08
OrdinalCLIP	$\mathbf{61.2 \pm 4.2}$	$\mathbf{0.47 \pm 0.06}$

Table 3: Results of different interpolation types with different numbers of base rank embeddings on the MORPH II dataset. Report MAE for each combination of settings.

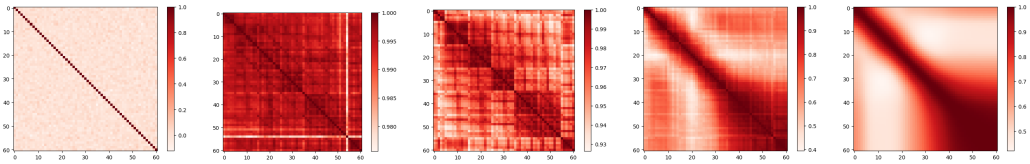
# Base Ranks	2	3	4	5	6	7	8	9
Linear	2.38	**2.32**	2.38	2.36	2.39	2.38	2.40	2.41
Inv. Prop	2.39	2.38	2.39	2.40	2.39	**2.35**	2.38	2.38

55,134 portraits from 13,618 individuals. Each portrait image is labeled with an age value from 16 to 77. Popular evaluation protocol [35, 39, 21, 20] is adopted in our experiments, which only contains 5,492 images of Caucasian descent to remove the cross-race interference. For general regression, we select 80% of the images for training, and others for testing. Note the long-tailed distribution in the dataset. Besides, the distribution of the test set constructed by random sampling is consistent with that of the training data. Adience contains 26,580 Flickr photos from 2,284 subjects. The label is annotated with eight groups, which are 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60-year-old. We adopt the same five-fold cross-validation protocol as used in [22]. We adopted mean average error (MAE) to measure the absolute differences between the ground truth labels and the predicted ones. In addition, classification accuracy is adopted for Adience. For the detailed experimental settings please refer to Appendix.

Comparison with State-of-the-art methods: We report the results in Table 1. The state-of-the-art methods are usually designed especially for age estimation [45, 41]. DRFs [39] learned a Gaussian mixture model by extending the deep neural decision forests [16] for regression. AVDL [45] leveraged a meta-learning framework to learn an adaptive variance for label distribution. POE [20] proposed a general regression framework with uncertainty modeling and achieved superior performance. Compared with the advanced approaches, zero-shot CLIP only obtains an MAE of 14.45, implying the poor modeling of numerical and ordinal concepts. CoOp scores an MAE of 2.39, illustrating the effectiveness of the language priors. OrdinalCLIP achieves a superior MAE of 2.32 without bells and whistles, resulting from the explicit ordinal constraint. We found similar results in the Adience [22] dataset. Specifically, CoOp attains 60.6% for regression accuracy, which is better than the previous best. Our OrdinalCLIP outperforms previous state-of-the-arts with an accuracy of 61.2% and is on par with the previous best method under the MAE metric of 0.47.

Parameter Analysis: We first analyze the influence of different numbers of base rank embeddings and different interpolation types. With all parameter combinations, the models can all achieve comparable MAEs on MORPH II(Table 3). When varying the number of base rank embeddings, the results improve slightly until a certain sweet point. The best MAE of 2.32 is reported with 3 base ranks. For *inverse-proportion* one, the best MAE of 2.35 is attained with 7 base ranks. However, its overall performance is sub-optimal compared with the *linear* interpolation one. In the following experiments, we chose *linear* interpolation and 3 base ranks as the default hyperparameters.

Ordinality of Learned Language Prototypes: We validate the ordinality of the rank embeddings by measuring the similarity of their corresponding language prototypes. OrdinalCLIP makes predictions by measuring the similarity between the normalized image embedding and each normalized language prototype in the CLIP latent space. Ideally, the structure of the normalized language prototypes should preserve the ordinality - "lie on a line" in the CLIP latent space. This is aligned with human perception and judgment while facing ordinal regression tasks. We compute the similarity matrices followed by Equation 4 with both inputs being language prototypes. The derived matrices are then



(a) CNN Baseline (b) CLIP, Rand. Ctx (c) CLIP, Init. Ctx (d) CoOp (e) OrdinalCLIP

Figure 3: The similarity matrices of language prototypes for different rank prompts. The redder, the more similar the pair of language prototypes. The percentages of prototype pairs that obey the ordinality are: 49.13%, 49.71%, 54.84%, 59.92%, and **65.94%**, respectively.

Table 4: Ablation experiments on the MORPH II dataset.

Choice	Regression Methods											
	CoOp [48]						OrdinalCLIP					
Tune Rank	✓		✓	✓		✓	✓		✓	✓		✓
Tune Ctx		✓	✓		✓	✓		✓	✓		✓	✓
Init Ctx				✓	✓	✓				✓	✓	✓
MAE	2.58	2.44	2.39	2.41	2.39	2.39	2.38	2.37	2.36	2.34	2.34	**2.32**

max-normalized by their maximum similarity. The ordinality score is the percentage of language prototype pairs that obey the ordinal property[2].

Figure 3 visualizes the ordinality under various situations. The CNN baseline (Figure 3(a)) has the least ordinality score of 49.13%; Half of the language prototype pairs violate the ordinal property and all the prototypes are distributed quite sparsely across the latent space. Figure 3(b) and Figure 3(c) are results computed directly from the CLIP model. Though the initialized context embeddings help to improve ordinality (from 49.71% to 54.84%), almost half of the language prototype pairs violate the ordinal properties. [33] found that CLIP models give degraded predictions facing abstract concepts like numbers. We attribute this to the loss of ordinality in the CLIP latent. Figure 3(d) shows that trained context embeddings from CoOp could further improve the ordinal property with an ordinality score of 59.92%. However, the learned language prototypes are still "twisted" in the CLIP space, leading to degraded predictions. Our OrdinalCLIP (Figure 3(e)) learns smoother language prototypes with a boosted ordinality score of 65.94%, which validates the effectiveness of the learned base rank and their interpolation. Note that the similarity matrices reflect the long-tail property of the dataset. Due to the insufficient samples for the older people, there is a "red blob" in the right-down corner. CLIP and CoOp cannot push the "older" labels away and leave them closer to the well-learned "younger" labels, which may lead to corrupted predictions. Our OrdinalCLIP reinforces the ordinality and as a result, outperforms the above two methods.

Ablation Study: To validate the effectiveness of OrdinalCLIP, we compare OrdinalCLIP with preceded CoOp [48] using different optimization and prompt context initialization strategies. There are four major observations in Table 4.

First, for both CoOP and OrdinalCLIP, solely optimizing either rank embeddings or context embeddings leads to sub-optimal performance, yet training them together results in optimal results. Note that CoOp only optimizes the context embeddings while leaving the rank embeddings untouched as they are initialized with the rank labels. However, after jointly training with the rank embeddings, the performance has been boosted from 2.44 to 2.39. For OrdinalCLIP, that is from 2.37 to 2.36 even with random base rank embeddings, resulting from the explicit ordinality from our method.

Second, the initialized rank embeddings are prone to be degraded, which is another observation validating that the CLIP latent space struggles to represent abstract concepts like numbers [33]. When only optimizing rank embeddings, CoOp obtains an MAE of 2.41 with initialized context prompts. In contrast, our OrdinalCLIP could achieve a better MAE of 2.38 even with all the context and base rank embeddings initialized randomly. We attribute this to the explicit ordinal modeling – the interpolation between the base ranks.

[2] *i.e.*, $(\sum_{i=0,j=0,i<=j}^{N-2,N-2} \mathbf{1}\{a'_{i,j} > a'_{i,j+1}\})/((N-1) \times N/2)$.

Table 5: We report the MAE results under few-shot settings on the MOPRH II dataset.

# Shots	1	2	4	8	16	32	64
CNN Baseline	8.04	6.00	4.66	3.91	3.57	3.06	2.84
CoOp [48]	5.09	4.50	3.81	3.57	3.23	2.87	2.61
OrdinalCLIP	4.94	4.36	3.55	3.31	3.07	2.76	2.57

Table 6: The MAE results under the distribution shift setting on the MOPRH II. "re cls" denotes the number of reduced classes, and "re smp" means the percentage of reduced sampled in one class.

re cls - re smp	10-80	10-90	20-80	20-90	30-80	30-90	40-80	40-90
CNN Baseline	2.88	2.98	3.46	3.68	3.19	4.06	3.28	3.65
CoOp [48]	2.71	2.85	2.98	3.51	3.06	3.36	2.99	3.30
OrdinalCLIP	2.61	2.67	2.77	3.06	2.86	3.21	2.84	3.12

Third, initializing prompts with task-specific context could boost performance, no matter whether it is CoOp or OrdinalCLIP, or which kind of embeddings are optimized. For example, our OrdinalCLIP could achieve the best MAE of 2.32 when training both embeddings and initializing the context.

Last, OrdinalCLIP outperforms CoOp under all settings in the ablation, which further validates the effectiveness of the proposed learned base rank embeddings and their interpolation.

Few-Shot Learning: We validate the generalization ability of OrdinalCLIP for the few-shot regression learning. The full dataset is split into 80% for training and 20% for testing. We use the entire test set for validation and only choose 1/2/4/8/16/32/64 samples in the training set from each class of the labels for training. We report the results in Table 5. We observe the language models consistently outperform the baseline by a large margin. By incorporating the language prior, the MAE is improved by 2.95 compared with baseline under 1 shot setting for CoOp; Our OrdinalCLIP further improves the MAE to 4.94. For other shot settings, we also obverse the same performance gains, which shows the effectiveness of the proposed learned rank embeddings.

Distribution Shift: Another generalization experiment about data distribution shifts is conducted on the MORPH II database. We also use the entire test set of the general regression setting. For the training set, we randomly choose several rank labels, and then randomly discard a large portion of samples in those classes. Table 6 shows the comparison results. We observe that by incorporating language prior, CoOp improves performance by a large margin compared with the baseline. While OrdinalCLIP boosts performance furthermore, illustrating our OrdinalCLIP can better resist the distribution shift of dataset by the explicit ordinal constraint. Specifically, with 40 rank labels selected and 90% of their samples discarded (most severe distribution shift setting), CoOp obtains a 0.35 improvement of MAE compared with baseline, while OrdinalCLIP further attains a performance gain of 0.18.

4.2 Image Aesthetics Assessment

Dataset. The image aesthetic dataset [37] contains 13,929 available Flickr photos of four categories, i.e, nature, animal, urban, and people. Five absolute rating scales are employed to evaluate the aesthetics quality: "unacceptable", "flawed", "ordinary", "professional", and "exceptional". Each image is judged by at least five different examiners and uses the median rank as ground truth. The whole dataset is randomly split into three non-overlapped subsets following [37]. Five-fold cross-validation was used for fair comparisons. We reported the mean values for both MAE and classification accuracy. We detail the experimental settings in Appendix.

Results: Table 7 shows the comparison with the previous SOTA methods. Language priors assist the model to achieve better performance, while zero-shot CLIP still struggles to distinguish ordinal concepts. Our OrdinalCLIP attains an overall MAE of 0.280 and overall accuracy of 72.58%, outperforming the previous best method. The results across all nominal categories either outperform or are on par with previous state-of-the-art. Consistent improvements are observed across all categories compared with baseline and CoOp, showing the effectiveness of learned rank prompts with the ordinal property.

Table 7: Quantitative Results on the Image Aesthetics dataset. We report accuracy and MAE.

Methods	Accuracy(%) - higher is better					MAE - lower is better				
	Nature	Animal	Urban	People	Overall	Nature	Animal	Urban	People	Overall
CNNPOR [22]	71.86	69.32	69.09	69.94	70.05	0.294	0.322	0.325	0.321	0.316
SORD [6]	73.59	70.29	**73.25**	70.59	72.03	0.271	0.308	**0.276**	0.309	0.290
POE [20]	73.62	71.14	72.78	72.22	72.44	**0.273**	0.299	0.281	0.293	0.287
CNN Baseline	72.76	70.44	70.12	70.80	71.03	0.338	0.381	0.364	0.377	0.365
Zero-shot CLIP [33]	63.33	36.78	49.84	24.79	43.69	0.438	0.737	0.683	1.063	0.730
CoOp [48]	72.74	71.46	72.14	69.34	71.42	0.285	0.298	0.294	0.330	0.302
OrdinalCLIP	**73.65**	**72.85**	73.20	**72.50**	**73.05**	**0.273**	**0.279**	0.277	**0.291**	**0.280**

4.3 Historical Image Dating

Dataset: The historical image dating dataset [26] is a benchmark for automatically predicting the decade of the historical colored image. There are five-decade categories in the database from the 1930s to the 1970s, where each category contains 265 images. For the general regression setting, the data of each category is divided into three parts: 210 for training, 5 for validation, and the rest for testing. Ten-fold cross-validation is adopted following [22, 26]. We report the mean and standard deviation for both MAE and accuracy metrics. Refer to Appendix for the detailed experimental settings.

Table 8: Results on the Historical Image Dating.

Methods	Acc (%)	MAE
CNNPOR [22]	50.12 ± 2.65	0.82 ± 0.05
GP-DNNOR [23]	46.60 ± 2.98	0.76 ± 0.05
POE [20]	54.68 ± 3.21	0.67 ± 0.04
CNN Baseline	42.56 ± 3.04	0.80 ± 0.03
Zero-shot CLIP [33]	26.08 ± 0.56	1.48 ± 0.03
CoOp [48]	51.90 ± 2.60	0.76 ± 0.06
OrdinalCLIP	$\mathbf{56.44 \pm 1.66}$	$\mathbf{0.67 \pm 0.03}$

Results: We compare our OrdinalCLIP with other state-of-the-art models using the full dataset. Table 8 demonstrates that CoOp improves the average accuracy by 9.34% compared with the baseline method, which again shows the importance of exploiting language priors. Zero-shot CLIP is still suffering from degraded results due to pool ordinal transferability. Our OrdinalCLIP further improves the accuracy by 4.45%. Compared with the previous best, OrdinalCLIP attains a competitive MAE of 0.67 and achieves a new state-of-the-art accuracy of 56.44% with only half of the standard deviation, which validates the effectiveness of the proposed learned rank embeddings.

5 Discussions and Conclusions

OrdinalCLIP is a general approach for ordinal regression. It can be deployed on many vision tasks such as age estimation. These tasks may be used for surveillance without user permission, which could lead to privacy issues. Therefore, strict technical controls may be needed. Since our method directly employs existing language models, the bias in the language models will be also inherited by our method. How to alleviate the bias in the language model still needs further research.

In this paper, we have presented the language-powered paradigm for ordinal regression. Existing ordinal regression methods usually suffer from the overfitting problem and lack of ordinality within the latents, as they mainly learn the rank concepts from the training data. We propose OrdinalCLIP, which associates each rank category with its language concept derived from the CLIP text encoder. To leverage the language priors, each rank is mapped to a corresponding language prototype. We further propose learnable rank prompts to explicit learning ordinal rank embeddings to preserve the order of the language prototypes in the language latent space. Extensive experimental results from three tasks including age estimation, historical image dating, and image aesthetics assessment show that our OrdinalCLIP attains very competitive performance in general ordinal regression tasks. OrdinalCLIP also gains improvements in few-shot and distribution shift settings for age estimation.

References

[1] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics (2017)

[2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)

[3] Chen, S., Zhang, C., Dong, M., Le, J., Rao, M.: Using ranking-cnn for age estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 (2017)

[4] Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada (2015)

[5] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers) (2019)

[6] Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (2019)

[7] Gao, B., Xing, C., Xie, C., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Trans. Image Process. (2017)

[8] Gao, B., Zhou, H., Wu, J., Geng, X.: Age estimation using expectation of label distribution learning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden (2018)

[9] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)

[10] Geng, X., Yin, C., Zhou, Z.: Facial age estimation by learning from label distributions. IEEE Trans. Pattern Anal. Mach. Intell. (2013)

[11] Gustafsson, F.K., Danelljan, M., Bhat, G., Schön, T.B.: Energy-based models for deep probabilistic regression. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX (2020)

[12] He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (2019)

[13] Jr., K.R., Tesafaye, T.: MORPH: A longitudinal image database of normal adult age-progression. In: Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK (2006)

[14] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (2017)

[15] Kong, S., Shen, X., Lin, Z.L., Mech, R., Fowlkes, C.C.: Photo aesthetics ranking network with attributes and content adaptation. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I (2016)

[16] Kontschieder, P., Fiterau, M., Criminisi, A., Bulò, S.R.: Deep neural decision forests. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015 (2015)

[17] Lathuilière, S., Mesejo, P., Alameda-Pineda, X., Horaud, R.: Deepgum: Learning deep robust regression with a gaussian-uniform mixture model. In: ECCV. pp. 202–217 (2018)

[18] Lee, J., Kim, C.: Image aesthetic assessment based on pairwise comparison A unified approach to score regression, binary classification, and personalization. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019 (2019)

[19] Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015 (2015)

[20] Li, W., Huang, X., Lu, J., Feng, J., Zhou, J.: Learning probabilistic ordinal embeddings for uncertainty-aware regression. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021 (2021)

[21] Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., Tian, Q.: Bridgenet: A continuity-aware probabilistic network for age estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (2019)

[22] Liu, Y., Kong, A.W., Goh, C.K.: A constrained deep neural network for ordinal regression. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)

[23] Liu, Y., Wang, F., Kong, A.W.: Probabilistic deep ordinal regression based on gaussian processes. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019 (2019)

[24] Martin, P., Doucet, A., Jurie, F.: Dating color images with ordinal classification. In: International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014 (2014)

[25] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013)

[26] Palermo, F., Hays, J., Efros, A.A.: Dating historical color images. In: Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI (2012)

[27] Pan, B., Wang, S., Jiang, Q.: Image aesthetic assessment assisted by attributes through adversarial learning. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019 (2019)

[28] Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)

[29] Pan, L., Ai, S., Ren, Y., Xu, Z.: Self-paced deep regression forests with consideration on underrepresented examples. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (2020)

[30] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021 (2021)

[31] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL (2014)

[32] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers) (2018)

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (2021)

[34] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning (2018)

[35] Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. (2018)

[36] Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)

[37] Schifanella, R., Redi, M., Aiello, L.M.: An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015 (2015)

[38] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers (2016)

[39] Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.L.: Deep regression forests for age estimation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)

[40] Shen, W., Zhao, K., Guo, Y., Yuille, A.L.: Label distribution learning forests. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (2017)

[41] Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., Li, S.Z.: Efficient group-n encoding and decoding for facial age estimation. IEEE Trans. Pattern Anal. Mach. Intell. (2018)

[42] Thawani, A., Pujara, J., Ilievski, F., Szekely, P.A.: Representing numbers in NLP: a survey and a vision. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021 (2021)

[43] Varamesh, A., Tuytelaars, T.: Mixture dense regression for object detection and human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020 (2020)

[44] Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)

[45] Wen, X., Li, B., Guo, H., Liu, Z., Hu, G., Tang, M., Wang, J.: Adaptive variance based label distribution learning for facial age estimation. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII (2020)

[46] Wu, H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2clip: Learning robust audio representations from CLIP. arXiv preprint arXiv:2110.11499 (2021)

[47] Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)

[48] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)

[49] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. arXiv preprint arXiv:2203.05557 (2022)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]